DOCUMENT RESUME

ED 207 982 SP 018 969

AUTHOR McBeath, Marcia; Carter, Delores
TITLE Staff Development Program Evaluation--Application of a Peer Supervision Program.
INSTITUTION District of Columbia Public Schools, Washington, D.C.
PUB DATE Aug 81
NOTE 17p.; Paper presented at the Annual Convention of the American Psychological Association (Los Angeles, CA, August 26, 1981).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; *Evaluation Methods; Inservice Teacher Education; *Lesson Observation Criteria; Observation; *Peer Evaluation; Program Evaluation; *Teacher Evaluation; *Teacher Supervision

ABSTRACT

This paper describes attempts to evaluate whether teachers who had participated in a staff development program could apply what they had learned in a controlled, specific situation significantly better than teachers who had not had the training. The program under evaluation provided training in peer supervision and evaluation. Test results covering knowledge of the techniques taught in the program showed significant gains by participants. Questionnaires and monitoring reports indicated that teachers did use the peer supervision methods learned through the project. An attempt is being made to develop and standardize an instrument to ascertain application of strategies taught in the course. A major revision of the evaluation procedure was in using a 16mm film instead of a videotaped lesson for the peer supervision participants to evaluate, with structured sections of a response form measuring three areas--observation, analyses, and conferences. Recommendations for future revision of the testing and evaluation procedures include: (1) pretraining staff members in scoring techniques; (2) requiring more specific statements from participants; and (3) measuring concurrent and predictive validity of the instrument from field observation. (JD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED207982

Staff Development Program Evaluation - Application of a Peer Supervision Program

Marcia McBeath

Delores Carter

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

✓ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. Cooper

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

SP 018 969

STAFF DEVELOPMENT PROGRAM EVALUATION - APPLICATION OF A PEER SUPERVISION PROGRAM[1]

Marcia McBeath[2] Delores Carter[3]

Introduction

It is difficult to evaluate the application of knowledge and techniques acquired in staff development courses after teachers have completed the training. Even though pre-post testing may indicate an increase in knowledge during the course, the question is, "Can the participants apply what they have learned one, two, and three years afterwards?" Frequently, the response on questionnaires indicates that they are using the material but (1.) self-reports are not always accurate, and (2.) participants may not be applying the knowledge correctly. This paper will describe one project's attempt to evaluate whether teachers who participated in a staff development program could apply what they had learned, in a controlled, specific situation, significantly better than teachers who had not had the training.

Funded by an ESEA Title IV-C grant in the District of Columbia Public Schools, a program has been in effect for four years to develop a resident instructional support system of peer supervision. Emphasis is placed on learning to apply the concepts of clinical supervision, a model developed by

---

(1) This paper was presented at the 1981 Annual Convention of the American Psychological Association, Los Angeles, California, August 26, 1981.

The project reported on in this paper was developed under a grant from ESEA Title IV-C in the Education Department. The contents do not necessarily represent the policy of the agency or endorsement by the Federal Government.

(2) Director, Title IV-C Evaluation Unit, Division of Research and Evaluation, District of Columbia Public Schools, Washington, D.C.

(3) ESEA Title IV-C Project Coordinator, IMPROVING INSTRUCTION THROUGH SUPERVISION: DEVELOPING SUPERVISORY SUPPORT FOR TEACHERS, Regions C and D, District of Columbia Public Schools, Washington, D.C.

SP 018969

Robert Goldhammer, which focuses on teachers assisting other teachers in their professional growth. The model uses a supportive, non-threatening, five-step approach consisting of a pre-observation conference, observation, analysis and strategy, post-observation conference and post-conference analysis. About 1000 teachers and administrators have been trained in the techniques through the project.

Of the 200 teachers who responded to a questionnaire asking for information on their use of the clinical supervision model during the 1979-80 school year, 83% indicated that they were using all or part of the model. Monitoring records of 311 teachers who were trained in the techniques showed that 73% used the model with fellow teachers during the first semester of the 1979-80 school year, and 54% during the second, serving 666 and 605 teachers respectively. The monitoring information also came from self-reports.

Original Instrument

In order to ascertain whether or not those who had had the training during any of the previous years when it was offered, could apply their knowledge in a given situation, an instrument was developed and administered to samples of teachers, some who had had the staff development training and some who had not. None of the tested participants, all curriculum leaders in the District, were told the purpose of the evaluation at the time. The instrument was administered during their monthly curriculum meeting. A standard introductory script began, "One of the focal points of a study in the school system is the observation of instruction. We are asking you to help in this study."

All participants were shown a ten-minute videotape of a teacher presenting a lesson. The standard script containing instructions was read to the group

by a researcher from the Title IV-C Evaluation Unit. The participants were to record their observations, analyze their recorded data and write what they would do next if they had the responsibility of providing instructional support for the videotaped teacher. Responses were recorded on a standard form. Criteria for scoring of the instrument were established by the project coordinator. Under "Observation of the Lesson," scorers were to count the number of factual, verbatim and judgmental statements. Under "Analysis of Data," they were to see if patterns in teaching behavior (style) were identified and if issues for post conference discussion were stated. Under "Next Step," they were to see if a post-observation conference was indicated, if self-analysis by the teacher was indicated, and if observer/teacher non-directive, collaborative interaction was indicated. Blind scoring was done by two impartial raters who had not taken the course. Inexperienced raters were used to see if written scoring criteria, alone, were adequate. Their returns, or observations, were recorded on a summary sheet. A third rating, was done by the project coordinator. In order to determine rater reliability in the scoring of the observation measures, criterion-related observer agreement of the raters was calculated. It was found to be from low to moderate.

No attempt was made during the first year to test the validity of the instrument except to examine it for face validity. Before being used with project participants and comparison groups, it was presented to team members, evaluation unit personnel, course instructors, and other project administrative personnel in order to increase the content validity. Using this feedback, some modifications were made before final use.

It was hypothesized that a sampling of teachers who had previously completed the training would demonstrate application of the training, in their analysis of a videotape lesson, to be measured by the differences of their responses from those of an equivalent sample who had not had the training, differences to be considered significant if they reached the .05 level on a t-test for two independent means.

Although the instrument was still in the design stage, results were analyzed on 94 participants, 25 who had received training and 69 who had not. There were statistically significant differences between the two groups in the predicted direction in three of the nine areas measured.

Second Year Changes

During the 1980-81 year of the project, refinements of the instrument were made and further standardization studies were carried out. The replication included the following changes:

1.) A film rather than a videotape was used for more clarity and convenience of presentation.

2.) Minor changes were made in the form used for recording the observation for more specificity.

3.) Raters were selected from personnel familiar with the model in an attempt to increase inter-rater reliability.

4.) Minor changes were made in the criteria for scoring the categories in an attempt to increase inter-rater reliability.

5.) Validity data were computed correlating film observation results with course knowledge test results and with a nationally standardized instrument.

The film used was part of a teacher training series showing a teacher working in a classroom situation.

The form used for recording the observation consisted of three parts. Part A in both the original and revised form, asked the participant to record his or her observations of the lesson.

Parts B and C were more specific. Part B said, "Review your recorded data. Indicate the behaviors that were most frequently recorded. Summarize the salient points in your recorded data."

The instruction for Part C were, "You have been asked to provide instructional support for the teacher in the film. Describe in detail your approach when you confer after the observation. Include some dialogue in your description of the conference."

Scoring fell into three general categories following the three areas on the observation form. Part A was "Observations," Part B was "Analyses" and Part C was "Conferences." These three areas were covered during course instruction and were included in the instructional manual. The general categories were broken down into sub-categories. Table I shows categories, sub-categories and examples.

TABLE I

Film Observation Peer Supervision
Instrument Scoring Examples

Category	Sub-Category	Example
Observations	Factual Statements	The students were asked to select the best place in their town for an incinerator.
	Verbatim Statements	"Let's pretend we are city workers and citizens. Where do you think we should place the town incinerator?"
	Non-verbal Statements	Teacher smiled after the student responded.
	Judgmental Statements	Some students were not interested in the small group activity.
Analyses	Summary Statements	Each student who wanted to was allowed to express his or her point of view.
	Teaching Patterns Identified	Teacher repeated each student's response.
Conferences	Indirect Approach	Teacher was asked how he felt the lesson went and what if any changes he would like to make in presenting that kind of lesson again.
	Direct Approach	Teacher was told, "I think you should have had more class discussion before the small group activity."
	Analysis Data Used as Guidelines	Teacher was asked if there was a reason for his repeating each student's response.
	Pre-Observation Data Recalled	Teacher was asked if he felt the objective of the lesson was accomplished.

The changes in the scoring criteria were mainly for clarity and better discrimination. The second version, for example, added non-verbal statements since awareness of recording such behavior had been added to the instructional manual.

Raters for the later version were all familiar with the project, being the project coordinator and course instructors. The project coordinator gave written instructions to the other raters on scoring criteria and indicated where the material was covered in the training guide. To determine inter-rater reliability Pearson Product Moment Correlations between two raters were calculated using an SPSS computer program. Table II shows the correlations.

TABLE II

Rater Reliability for the Film Observation Peer Supervision Instrument

OBSERVATIONS				ANALYSES		CONFERENCES			
Factual Statements	Verbatim Statements	Non-verbal Statements	Judgmental Statements	Summary Statements	Teaching patterns	Indirect Approach	Direct Approach	Used Analysis Data	Recalled Pre-Observation Data
.87	.99	.68	.83	.51	.87	.84	.12	.23	.31
N=84	N=84	N=84	N=84	N=82	N=82	N=60	N=61	N=59	N=59

Although the rater reliability is not high in several areas, the project coordinator felt that with revision of scoring criteria and training of raters, higher agreement can be reached.

To compute concurrent validity data the scoring of the Film Observation Peer Supervision Instrument by the project coordinator was used for all subjects. The first instrument used was a project designed knowledge test that has demonstrated measurement of a significant gain pre- to posttesting as well as significant differences between experimental and control groups. The second measure was a nationally standardized test, the supervision and administration section of the National Teachers Examination (NTE). Results of the Pearson Product Moment Correlation computed using the SPSS package are shown in Table III.

TABLE III

Concurrent Validity of the Film Observation Peer Supervision Instrument

OBSERVATIONS				ANALYSES		CONFERENCES			
Factual Statements	Verbatim Statements	Non-verbal Statements	Judgmental Statements	Summary Statements	Teaching patterns	Indirect Approach	Direct Approach	Used Analysis Data	Recalled Pre-Observation Data
* .25	.17	.19	.06	.22	.11	.21	-.21	-.02	.05
N=56	N=55	N=53	N=54	N=56	N=55	N=51	N=41	N=50	N=50
** .09	.26	.25	.14	.18	.34	-.03	-.02	.07	.08
N=34	N=33	N=31	N=33	N=34	N=33	N=30	N=24	N=29	N=29

* With Project Designed Supervision Test

** With NTE

As may be seen in Table III, there is some positive correlation with at least one of the tests in all areas except the Direct Approach. Since this area also had low reliability ratings, it will be dropped from the next revision.

No attempt has been made to determine a composite score which might yield higher validity. Correlation with a real observation of participants using the techniques learned should be done concurrently and predictively since the knowledge test and the NTE measure different aspects of peer supervision. However, this has been impractical up to the present. The reported scores do add some information for development of this instrument.

Results of Use

With a district reorganization in 1980-81, six regions were changed to four, bringing new schools into the regions where the peer supervision project began. An experimental design was set up where 10 of the new schools were randomly assigned to receive the training and 10 to a control group. The control group were invited to participate in the training during the 1981-82 school year. The 42 control group teachers who were tested were those who were interested in enrolling in the course. There were 74 teachers in the experimental group. Pre-post test results of the Film Observation Peer Supervision Instrument are shown in Tables IV and V. A t-test for related measures was done on the first six items of the instrument and a chi-square on the last four (the yes/no items). All computation was done using SPSS packages.

TABLE IV

Pre-Post Test Results on the Film Observation
Peer Supervision Instrument for Items 1-6

	N=49 Factual Statements	N=49 Verbatim Statements	N=48 Non-Verbal Statements	N=48 Judgmental Statements	N=48 Summary Statements	N=48 Teaching Patterns
Pretest Mean	5.61	0.00	0.04	1.71	3.38	0.31
Posttest Mean	10.04	6.61	0.69	0.83	7.00	1.56
Pretest SD	3.75	0.00	0.20	1.75	2.28	0.93
Posttest SD	5.52	8.22	1.13	1.62	3.53	1.43
T-Value	-5.89	-5.63	-3.99	2.93	-7.02	-4.90
2-tested probability	.000	.000	.000	.005	.000	.000

TABLE V

Pre-Post Test Results of the Film Observation Peer Supervision Instrument for Items 7-10

	N=39(PRE) N=40(PO) Indirect Approach		N=40(PRE) N=33(PO) Direct Approach		N=39(PRE) N=39(PO) Used Analysis Data		N=39(PRE) N=39(PO) Recalled Pre. Obs. Data	
	PRE	POST	PRE	POST	PRE	POST	PRE	POST
Yes (Number)	11	28	33	7	16	23	5	34
Yes (Percent)	28.2	71.8	82.5	17.5	41.0	59.0	12.8	87.2
No (Number)	32	8	9	24	33	6	35	4
No (Percent)	80.0	20.0	27.3	72.7	84.6	15.4	89.7	10.3
Chi Square	19.32		20.27		14.05		43.16	
Significance	.000		.000		.000		.000	

As may be seen in Tables IV and V, there is a significant increase from pre- to post-training in peer supervision in all the areas measured by this instrument.

Experimental-Control Group results are shown in Tables VI and VII. A t-test for independent means was done on the first six items of the instrument and a chi-square on the last four (the yes/no items). All computations were done using SPSS packages.

TABLE VI

Experimental- Control Results on the Film Observation Peer Supervision Instrument for Items 1-6

	N=56(EXP) N=42(CON) Factual Statements	N=55(EXP) N=42(CON) Verbatim Statements	N=53(EXP) N=43(CON) Non-Verbal Statements	N=54(EXP) N=42(CON) Judgmental Statements	N=56(EXP) N=42(CON) Summary Statements	N=55(EXP) N=41(CON) Teaching Patterns
Exper. Mean	9.93	6.09	0.62	0.80	6.75	1.47
Control Mean	6.12	0.14	0.07	2.90	4.12	0.17
Exper. SD	5.59	7.93	1.10	1.55	3.49	1.39
Control SD	4.34	0.52	0.34	2.92	2.58	0.44
T-Value	3.80	5.55	3.46	-4.24	4.29	6.54
2-Tested Probability	.000	.000	.001	.000	.000	.000

TABLE VII

Experimental- Control Results on the Film Observation Peer Supervision Instrument for Items 7-10

	N=51 (EXP) N=42(CON) Indirect Approach		N=41(EXP) N=42(CON) Direct Approach		N=50(EXP) N=42(CON) Used Analysis Data		N=56(EXP) N=42(CON) Recalled Pre. Obs. Data	
	EXP	CON	EXP	CON	EXP	CON	EXP	CON
Yes (Number)	39	7	31	21	41	7	42	7
Yes (Percent)	76.5	25.0	31.7	75.0	82.0	25.0	84.0	25.0
No (Number)	12	21	28	7	9	21	8	21
No (Percent)	23.5	75.0	68.3	25.0	18.0	75.0	16.0	75.0
Chi Square	17.63		10.80		22.29		24.28	
Significance	.000		.000		.000		.000	

As may been seen in Tables VI and VII, there is a significant difference between the experimental and control groups in all areas measured by this instrument.

Summary and Discussion

A Title IV-C project in Regions C and D in the District of Columbia, has been training teachers and administrators in peer supervision techniques for the past four years. Test results covering knowledge of the techniques have consistently shown significant gains made by participants. Questionnaires and monitoring reports indicate that teachers are using the peer supervision methods learned through the project. However, these are all self-reports and they do not show that teachers are applying the techniques correctly.

So far, it has been impractical to observe and rate teachers in real supervising situations although role playing is one of the techniques used in the course.

Presently an attempt is being made to develop and standardize an instrument to ascertain application of the strategies taught in the course. This was first attempted during the 1979-80 school year using a videotape lesson for the peer supervision participants to react to as they would a real observation, with structured sections of a response form developed to measure the three areas - observations, analyses and conferences.

Some inter-rater reliability and test use studies were done leading to the first revision. One major change was the use of 16 M.M. film instead of a videotape. In the present study there were four sub-sections scored under Observations and inter-rater reliabilities (Pearson Product Moment) ranged from .68 to .99. In the two sections under Analysis the range was .51 to

.87. In the four sections under Conferences, the range was .12 to .31

Except for the project coordinator, none of the raters had previously scored the instrument. As course instructors they, however, were familiar with the techniques. For their ratings they used written descriptions of criteria and references to the training manual. The next step will be training raters in advance to try to produce greater reliability. Some instructions may also need revision.

Correlations were also made to begin validity measures. The Pearson Product Moment for the four sub-tests under Observations ranged from .06 to .25 with the project designed peer supervision knowledge test and from .09 to .26 on the supervision/administration section of the NTE. Correlations on the Analysis section ranged from .11 to .22 on the project test and from .18 to .34 on the NTE. On the Conferences section, correlations ranged from -.21 to .21 on the project test and from -.02 to .08 on the NTE.

Although the validity correlations were low, it was known that neither test measured exactly the same material sought in the film observation instrument. A better method of validity would be correlating results of the instrument with field observations for concurrent validity and with future observations for predictive validity. It may be possible to do some of this in the future.

The instrument was used for pre-and posttesting the experimental group. It was also used in comparing a control and an experimental group. In the

pre-post testing, statistically significant gains were found in all sections including a decrease in judgmental statements as hypothesized. In the experimental-control comparisons, similar statistically significant gains were found in all sections.

Recommendations for future revisions of the instrument include the following:

1. Participants' record sheet should require more specific responses such as "State three questions you might ask the teacher during a conference."
2. The category, Direct Approach, should be omitted since reliability and validity data are so low.
3. Inter-rater reliability should be improved by training raters and revising criteria if indicated.
4. Concurrent and predictive validity of this instrument should be measured using field observation ratings.

Even in its present state, however, the instrument has been useful in demonstrating that the application of techniques learned in staff development programs can be measured using a group test.